DOCUMENT RESUME

ED 414 748 FL 024 943

AUTHOR Pearson, Jennifer
TITLE Strategies for Identifying Terms in Specialised Texts.
PUB DATE 1996-00-00
NOTE 11p.; For serial publication in which this article appears, see FL 024 940.
PUB TYPE Journal Articles (080) -- Reports - Descriptive (141)
JOURNAL CIT TEANGA: The Irish Yearbook of Applied Linguistics; n16 p33-42 1996
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Discourse Analysis; Foreign Countries; *Language Patterns; Language Research; *Languages for Special Purposes; *Lexicography; Linguistic Theory; *Vocabulary
IDENTIFIERS Language Corpora

ABSTRACT

A study of two language corpora (recommendations and specifications produced by the International Telecommunications Union for its members, and a corpus containing university syllabi for economics, biology, and history) investigated lexicographic strategies for identifying terms in specialized texts. In the first phase, a list of all possible term formations was compiled by examining the composition of words and phrases known to be terms, and their patterns were analyzed. In the course of this analysis, a number of signals of the presence of a term within a sentence were noted, and these were used to refine the term identification process, first by identifying all term candidates and many non-terms in the corpora, and then by selecting those term candidates that also satisfied the generic reference criteria (i.e., it must be unmarked and not modified by a word that does not form an integral part of the term) and at least one additional criterion based on the occurrence of the signals. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 414 748

Strategies for identifying terms in specialised texts

Jennifer Pearson
Dublin City University

Introduction

The compilation of specialised glossaries is frequently an integral function of translation departments involved in the translation of specialised texts. Glossary compilation is a time-consuming undertaking whereby terminographers have to sift through large bodies of text in order to identify and retrieve terms and their equivalents in other languages. In the past and perhaps even still, much of this work was done manually.

The first step in the glossary compilation process is term identification. Terms are those words or phrases which are considered to form part of the special vocabulary of a particular subject domain. As the many articles devoted to the topic will testify, term identification is a lot less straightforward than one might imagine. The terminographer has to decide whether certain words or phrases should be considered as having special reference, and therefore terminological status, within a particular context or whether it is more sensible to consider them as part of general vocabulary. S/he also has to decide where the cut-off point for a multi-word term should be. For example, should the telecommunications term *amplitude companded single sideband modulation* be classified as a single term or as two or more separate terms? The terminographer generally relies on a combination of specific criteria and his/her own intuition to allow or eliminate term candidates. Specific criteria for inclusion will generally include the relative frequency of a word or phrase in a text.

Automatic identification of terms

As mentioned above, the automatic identification of terms has already engaged the minds of a number of researchers, primarily those working in information retrieval and natural language processing. The identification of terms in telecommunications texts is well documented, by Béatrice Daille (1994) in her PhD thesis on the topic, by researchers at Dublin City University and UMIST as part of the EU funded Eurotra research project.

EL024943

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Jeffrey Kallen
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Yang (1986) has devised a technique for identifying scientific and technical terms in a scientific English corpus. At the University of Surrey, work on terminology retrieval has been undertaken in a number of fields including automotive engineering and aeronautics. Jacquemin and Royaute (1994) have worked with the MEDIC corpus which is a bibliographical medical corpus. As these examples show, the research tends to be subject field specific which is not unreasonable given that term formation patterns may vary from one subject field to another.

Research has generally focused on examining term formation patterns which occur in corpora with a view to tagging the corpora and retrieving term candidates. Tagging involves assigning a predetermined set of word classes to all words in a corpus. Each word is identified as belonging to a certain part of speech. A number of different tag sets exists, and this author has used the tag set devised by the Corpus Linguistics Group at the University of Birmingham. In a preliminary phase, a manual analysis is carried out to identify the composition of terms in a corpus and a list is drawn up of all possible combinations. These combinations are then used as input to retrieve all term candidates. Nkwenti-Azeh (1992) attempted to identify potential terminological units using a positional and combinational approach. He found that:

> the positional approach .. removes the need to comprehensively mark up the terms of an input text, especially where we are dealing with a circumscribed corpus: terminological units occurring in the text can be reconstituted if the positions of their elements have been specified in the positional database. (1992:19)

Jacquemin and Royaute (1994) were in fact more interested in retrieving term variants than actual terms. They used an existing set of terms and an analysis of the head-modifier relations to establish whether other syntactic patterns containing the same head and modifier could be classified as a variant of a term contained in their list.

Yang used frequency and collocational patterns to identify terms. He states:

> Since terms are highly subject matter specific, it is possible to identify single-worded terms on the basis of their frequencies of occurrence and distribution. Multi-worded terms are identified on the basis of their collocational behaviour. (1986:93)

The problem with the criterion of frequency is that it excludes terms which occur only once or infrequently in a corpus. Daille (1994) combined morpho-syntactic and statistical approaches to extract term candidates. She focused exclusively on binary term formation patterns (e.g. adj+noun) in order to write a program for retrieving all such patterns from her corpus. Frequency was once again an important criterion for assessing the eligibility of a term candidate.

There is one not inconsiderable drawback to the term formation approaches outlined above. While the specification of term formation patterns will ensure that all candidates meeting the required specification will be considered as potential terms, it will also ensure the inclusion of many words or phrases which will prove to be non-terms. For example, if the pattern *adj+noun* has been specified as a term pattern, all occurrences of modified nouns, regardless of their status, will be included, resulting in the retrieval of a far greater set of potential terms than one would wish to have. For this reason, I decided to explore the possibility of applying an additional set of criteria which would refine the selection process and result in a more accurate list of term candidates.

The corpora which I used for my analysis are the ITU corpus (4.7m words) and the GCSE corpus (1m words). The ITU corpus is available on CD-ROM from the University of Edinburgh and comprises recommendations and specifications produced by the International Telecommunications Union for its members. The GCSE corpus, to which I have access via the COBUILD unit at the University of Birmingham, contains some of the GCSE syllabus for economics, biology and history.

Identifying term formation patterns

In the first phase, I produced a list of all possible term formation patterns in each of the corpora by examining the composition of words/phrases which I knew to be terms. I considered them to be terms because they co-occurred with certain hinges such as *denotes, is defined as, is called.* As term formation patterns tend to vary from one text type to another (i.e. complex terms are more common in texts where the author-reader relationship is one of expert to initiated reader), I decided that it would be more useful to analyse term formation patterns in each of the corpora separately. The corpora were tagged using the CLG tagger devised by the Corpus Linguistics Group at the University of Birmingham.

Term formation patterns in the ITU corpus

The list of term formation patterns in Table 1 contains those which were identified by examining all nouns and noun phrases which occurred with connective verbs such as *denotes, is defined as, is called.*

Table 1. Term formation patterns, ITU corpus

Term formation pattern					Term candidate
+det	NN				a window
					a pointer
+det	JJ	NN			an effective call
+det	NN	NN			a test cycle
					the message group
+det	JJ	NN	NN		a generic test suite
					a compound parameter name
+det	JJ	NN	NNS		the low layer capabilities
+det	NN	IN	NN		the plane of measurement
+det	NN	NN	NN		the frame alignment procedure
					the TCAP message format
					a circuit multiplication system
+det	VBN	NN			a confirmed service
					the applied load
+det	VBG	NN			the magnifying optics
+det	NN	VBN	NN		A position defined parameter
+det	VB	NN	NN		The resynchronize type parameter
					The receive state variable
+det	NN	NN	VBG	NN	The data link resetting procedure[1]
-det	NN				interworking
-det	NN	NN			envelope delay
-det	NN	NN	NN		access control administration
-det	JJ	NN	NN		direct parameter input
-det	JJ	NN	NNS		analogue transfer links
-det	JJ	NN			arithmetic subtraction
					dynamic alignment
-det	JJ	NNS			functional entities
-det	NN	JJ			implementation dependent
-det	VBN	VBG			compelled signalling

As the examples show, terms can have up to four components but such terms will be rare and, in many instances, what appear to be 4-component terms are in reality either modified 3-component terms, or terms which have a generic class word such as *system, procedure* or *method* as the head word of the NP.

At first glance, a considerable number of the terms do not strike one as being particularly 'technical'. These include terms such as *window, effective call, offered load, applied load.* If a terminologist were to rely on her/his own intuition, s/he might not identify these as term candidates. However, we know that they are terms because the phrases with which they occur function as hinges to indicate the presence of terms. Moreover, subject experts have confirmed that these term candidates are indeed terms.

Term formation patterns in the GCSE corpus

Table 2. Term formation patterns, GCSE Corpus

Term formation patterns	Term candidates	Term formation patterns	Term candidates
+ det VBN NN	a balanced diet	!+det NN	a balk
	a closed system		a barometer
	a concealed seam		a calorie
	a sex-linked disease		a carrier
- det NN NN	bank cashier		a caterpillar
- det NN NNS	linotype operators	-det NN	chalk
+ det VBG NN NN	a scanning electron micrograph		chloramphenicol
			coal
			barley
		+ det JJ NN	a binary system
			a double circulation
			a dry cell
			a wet pit
			a white dwarf
		- det JJ NN	hydrochloric acid
			electric drills

The list of term formation patterns in Table 2 contains those which were identified by examining all nouns and noun phrases which occurred with hinges such as *called* and *e.g.* Unlike the terms in the ITU corpus which were frequently complex or multi-word terms consisting of 3 words, the terms in the GCSE corpus are generally single word or 2 word terms. There is a very small number of 3-word terms.

Refining the term identification process

I suggested previously that it might be useful to consider applying a further set of criteria in order to refine the term identification process. The purpose of the additional set of criteria would be to prevent certain items

from being considered as term candidates, even when they meet the corpus-specific term formation criteria which would normally allow them to be considered.

In analysing descriptive patterns in the corpora, I observed a number of hinges which signalled the presence of a term in a sentence. I decided to use these hinges in order to refine the process of term retrieval. I envisaged that the process would occur in two stages. The first part of the process involves an analysis of the corpora using the corpus-specific term formation criteria, as outlined previously. This identifies all term candidates and many non-terms in each of the corpora. The second part of the process selects only those term candidates which also satisfy the generic criterion (cf below) and **at least one** of the additional criteria which are based on the occurrence of the hinges described below.

While it was necessary, for the first part of the term identification process, to devise a set of corpus-specific term formation patterns, this section describes specifications which can be applied to all corpora which have an informative function.

The first and, I believe the most important criterion is that of generic reference. Generic reference is one of the key tenets of the traditional theory of terminology where a clear line is drawn between generic concepts and individual objects.

It should always be borne in mind that concepts cannot be taken for the individual object themselves. They are mental constructions serving to classify the individual objects of the inner or outer world by way of a more or less arbitrary abstraction. ISOR 704 Naming Principles (1968:8).

ISO makes a distinction between the generic concept and the individual object, i.e. the realization of that concept by virtue of its location in time and space. Picht/Draskau (1984) also make a distinction between the generic and the individual. They prefer, however, to distinguish between a generic concept and an individual concept rather than an individual object. Like ISO, they argue that the presence or absence of definiteness, i.e. whether or not the concept can be located in time and space will determine whether or not the term is to be considered generic or individual.

An individual concept will be represented by a *name* rather than a *term*. The notion that the absence or presence of an indication of time and space allows us to distinguish between the generic and the individual respectively is an interesting one. However, while the theoretical approach to terminology

makes this distinction, it does not make any reference to how this distinction is realized in text, and this is what is of particular interest to us.

We have noted in our corpora that terms are referred to in two ways. Broadly speaking, terms are either marked or unmarked. When I use the expression *marked term*, I mean that it may be preceded by any number of determiners, with the exception of the indefinite article. When I use the expression *unmarked term*, I mean that the term is preceded by the indefinite article or is not preceded by any article at all. When a term is marked, the reference is likely to be specific. The author is situating the use of the term in time and space, in this instance within a particular text. When unmarked, the reference is likely to be generic. I have ascertained, for example, that when a term is not marked by any indication of definiteness, it refers to its generic concept. When the same term is marked by definiteness or when its environment does not meet the required conditions, it is not possible to assume that reference is being made to the generic concept. There are, however, exceptions to this rule. For example, there are instances where a marked term at the beginning of one sentence functions as an anaphoric reference to a generic reference in the previous sentence. However, we need not concern ourselves with these for identification purposes because the generic reference will already have been identified.

The first condition which term candidates must meet is: *A term candidate must have generic reference*. For a term to have generic reference, it must be *unmarked*, i.e. preceded by the indefinite article or by no article at all.

The second condition which I specified relates to modifiers. If a term candidate is modified by a word other than a word which is considered to be part of the term itself, it is considered to have specific rather than generic reference. The second condition, therefore, is: *A term candidate may not be modified by a word which does not form an integral part of the term*.

If a term candidate satisfies each of these two criteria and occurs with one of the hinges described below, it may be considered to be a term.

The hinges which I have identified in the two corpora as signalling the presence of a term are: *called, known as, e.g., the term, is/are defined as, denote/denotes, consist/consists of, comprise*. The combination of the term formation selection restrictions with the generic and hinge restrictions appears to be sufficient for the purposes of term identification. Examples of occurrences of some of **the hinges are provided below.**

Examples with *called* from the ITU corpus

ate-oriented version - The state is described uniquely using pictorial elements. This picture is **called** a state picture. - The transition action sequence is implied by the difference between

generic name for a set of applications provided to telematic users. Each of those applications is **called** a telematic interworking application (TIA). Access to and participating in

present the DTE identity in the address fields of#IT# call set-up #IQ# packets. This number is **called** a "DTE address" and is defined in § 3.1.3. #TITRE# 2.3.1 #IT# Service attributes

values is an integer which references an instance of use of an abstract syntax. The integer is **called** a #GR# presentation context identifier #AS# and fills the "indirect-reference INTEGER" field

Examples with *called* from the GCSE corpus

< one by a narrow piece of land **called** a balk. Why was this bad farming?)
< can be measured by an instrument **called** a barometer. There are two main >
< pulse of electricity. This code is **called** a binary system. The pulses travel >
< star and the two stars together are **called** a binary star. The Solar System has >

Examples with *known as* from the ITU corpus

rk are identified by a unique code **known as** a point code (Recommendation Q.70
g routes in their priority order is **known as** a signalling route set. One signall
ly released telephone connection is **known as** a cutoff call when the connection
the order of 100 to 1000 metres is **known as** an extended passive bus. This confi

Examples with *known as* from the GCSE corpus

< directly to the electorate is **known as** a referendum. Holding a >
< through the air as a cloud. This is **known as** a POWDER AVALANCHE. >
< The Palestine Arab guerrilla force, **known as** Al Fatah, operated against >
< the amoeba to help it move. (This is **known as** amoeboid motion # An amoeba >

Examples with *e.g.* from the ITU corpus

ion are provided by different media (**e.g.** a satellite channel in one direction
whom the charges are to be charged (**e.g.** a branch, a bank or a similar institu
it/s channel time slot can accomodate **e.g.** ,a PCM-encoded voiceband signal co
rough another communication system (**e.g.** a physical delivery system) that is

Examples with *e.g.* from the GCSE corpus

following groups:Class I Professional **e.g.** chartered accountants, senior civil >
< drugs are produced in laboratories, **e.g.**chloramphenicol # used to combat >
< years) and are called non-renewable, **e.g.** coal. problems will be caused if any
< Some food is dried in a vacuum, **e.g.**coffee granules. This is called >

Examples with *The term* from the ITU corpus

alignment #IQ# For the 6312 kbit/s hierarchical level, **the term** "frame alignment" is synonymous with "multiframe align
u¡h which the desired analogue points could be derived. **The term** "virtual analogue switching points" is alşo used for "vice primitives", "peer protocol" and "peer entities". **The term** "boundary" applies to boundaries between layers telex subscriber over the telex network. #TITRE# 12.6 **The term** "notification" applies to the forwarding of an advice

Examples with *The term* from the GCSE corpus

< energy is passed on to animals. **The term** biomass is used to refer to anything
< referred to as bureaucrats. **The terms** bureaucracy' and 'bureaucrats' have
< are not the Coastal management **The term** COASTAL MANAGEMENT is a <
camps of the Boer War **The term** 'concentration camp derives from the

Conclusion

When embarking on this research, I started from the premise that previous efforts to identify and retrieve terms from a large body of text had resulted in the inclusion of many words or phrases which could not be considered as terms in the context. I used the same type of positional approaches adopted by researchers such as Daille and Nkwenti-Azeh but added a set of selectional syntactic restrictions to refine the term identification process. I have shown that it is important for a term candidate to have generic rather than specific reference and that the generic criterion can be formulated and implemented in the retrieval process. As the corpora which I was using had an informative function, I was able to use the information signalling hinges in my selection restrictions. I believe that it should be possible in other text types to avail of other hinges (e.g. certain types of verbs, position in sentence) to refine the term retrieval process. Future research will include an investigation of other text types to assess the general applicability of the approach proposed here.

Footnote

[1] This is an example of an NP with a generic class word as head of the NP. I would envisage stripping the NP of the head word at a later selection stage.

Bibliography

Daille, B. (1994). Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques. PhD Thesis. Université Paris VII.

ISO R704 (1968). Naming Principles.

Jacquemin, C., Royaute, J. (1994). Retrieving Terms and their Variants in a Lexicalized Unification-Based Framework in *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval.* Springer-Verlag. pp. 132-141.

Nkwenti-Azeh, B. (1992). Positional and Combinational Characteristics of Satellite Communications Terms. *Eurotra-UK Final Report.*

Picht, H., Draskau, J. (1984). *Terminology: an Introduction.* UNESCO, Paris.

Yang, H.Z. (1986). A new technique for identifying scientific and technical terms and describing science texts in *Literary and Linguistic Computing,* Vol. 1, No.14, pp 93-103.

11